Under-Determined Reverberant Audio Source Separation Using Local Observed Covariance and Auditory-Motivated Time-Frequency Representation

Ngoc Q. K. Duong, Emmanuel Vincent and Rémi Gribonval

INRIA, Centre Inria Rennes - Bretagne Atlantique, France
qduong@irisa.fr; emmanuel.vincent@inria.fr; remi.gribonval@inria.fr

Abstract. We consider the local Gaussian modeling framework for under-determined convolutive audio source separation, where the spatial image of each source is modeled as a zero-mean Gaussian variable with full-rank time- and frequency-dependent covariance. We investigate two methods to improve the accuracy of parameter estimation, based on the use of local observed covariance and auditory-motivated time-frequency representation. We derive an iterative expectation-maximization (EM) algorithm with a suitable initialization scheme. Experimental results over stereo synthetic reverberant mixtures of speech show the effectiveness of the proposed methods.

Key words: under-determined convolutive source separation, nonuniform time-frequency representation, ERB transform, full-rank spatial covariance model

1 Introduction

Multichannel under-determined source separation is the problem of recovering from a vector of I observed mixture channels $\mathbf{x}(t)$ either the J source signals $s_j(t)$ or their *spatial images* $\mathbf{c}_j(t)$, *i.e.* the vector of their contributions to all mixture channels, with $1 < I < J$. The mixture of several sound sources, *e.g.* musical instruments, speakers and background noise, recorded by the microphone array is then expressed as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t). \tag{1}$$

We consider a reverberant environment where $\mathbf{c}_j(t)$ can be modeled via the convolutive mixing process

$$\mathbf{c}_j(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t-\tau) \tag{2}$$

where $\mathbf{h}_j(\tau)$ is the vector of filter coefficients modeling the acoustic path from source j to all microphones.

Most existing approaches, *e.g.* [1], first transform the signals into the time-frequency (T-F) domain via the short time Fourier transform (STFT) and approximate the convolutive mixing process (2) under a narrowband assumption by complex-valued multiplication in each frequency bin f and time frame n

$$\mathbf{c}_j(n,f) \approx \mathbf{h}_j(f) s_j(n,f) \tag{3}$$

where $s_j(n,f)$ and $\mathbf{c}_j(n,f)$ are the STFT coefficients of $s_j(t)$ and $\mathbf{c}_j(t)$, respectively, and the mixing vector $\mathbf{h}_j(f)$ is the Fourier transform of $\mathbf{h}_j(\tau)$. In [2,3] we considered a different framework where $\mathbf{c}_j(n,f)$ is modeled as a zero-mean Gaussian vector random variable with covariance matrix

$$\mathbf{R}_{\mathbf{c}_j}(n,f) = v_j(n,f)\,\mathbf{R}_j(f) \tag{4}$$

where $v_j(n,f)$ are scalar time-varying *variances* encoding the spectro-temporal power of the sources and $\mathbf{R}_j(f)$ are time-invariant *spatial covariance matrices* encoding their spatial position and spatial spread. We proposed to represent $\mathbf{R}_j(f)$ as a full-rank matrix and showed that this better models reverberation and improves separation compared to the rank-1 spatial covariance model resulting from the narrowband assumption, *i.e.* $\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f)$ with ^{H} denoting matrix conjugate transposition [3]. Nevertheless, the *model parameters* $v_j(n,f)$ and $\mathbf{R}_j(f)$ are often inaccurately estimated due to T-F overlap between sources. In this paper, we present two complementary methods to further enhance separation performance by respectively improving the robustness of parameter estimation to source overlap and reducing this overlap.

Firstly, instead of estimating the model parameters from the mixture STFT coefficients $\mathbf{x}(n,f)$ as in [3], we infer them from the observed mixture covariance $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$ in the neighborhood of each T-F bin in a maximum likelihood (ML) sense. This method was introduced in [4] in the context of instantaneous audio source separation as a sliding window variant of the T-F patch-based model in [5]. Besides the interchannel phase and intensity differences encoded by $\mathbf{x}(n,f)$, $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$ also encodes correlation between the mixture channels which decreases with the larger number of active sources and the larger angle between these sources. This additional information results in improved separation of instantaneous mixtures: for instance, time-frequency bins involving a single active source are perfectly estimated since they are characterized by rank-1 local observed covariance *i.e.* maximum interchannel correlation [4, 6]. In the following, we show that this method also improves separation performance on reverberant mixtures, despite the fact that interchannel correlation is intrinsically lower in this context.

Secondly, we investigate the use of a nonuniform T-F representation on the auditory-motivated equivalent rectangular bandwidth (ERB) scale. ERB-scale representations have been widely used in computational auditory scene analysis (CASA) [7] and in a few other studies [8,9] since they provide finer spectral resolution than the STFT at low frequencies, hence decrease the overlap between sources in this crucial frequency range where most sound energy lies. These representations were shown to improve separation performance based on ℓ_1-norm minimization for instantaneous mixtures [9] or single-channel Wiener filtering for convolutive mixtures [8]. Yet, they failed so far to improve separation performance based on more powerful multichannel filtering techniques for reverberant mixtures, due to the fact that the narrowband approximation does not hold because of their coarse spectral resolution at high frequencies [8]. In the following, we show that ERB-scale representations are also beneficial for multichannel convolutive source separation provided that the full-rank covariance model is used.

The structure of the rest of the paper is as follows. We define a ML criterion exploiting local observed covariance and applicable to auditory-motivated T-F representations in Section 2. We then derive an associated source separation algorithm in Section 3 and report some experimental results in Section 4. Finally we conclude in Section 5.

2 ML criterion exploiting local observed covariance and auditory-motivated T-F representation

We first derive the likelihood using the local observed covariance in the neighborhood of each STFT bin. We then extend this criterion to an auditory-motivated T-F representation resulting from the ERB scale.

2.1 ML inference employing local observed covariance

Let us assume that the source image STFT coefficients $\mathbf{c}_j(n', f')$ follow a zero-mean Gaussian distribution over some neighborhood of (n, f) with covariance matrix $\mathbf{R}_{\mathbf{c}_j}(n, f)$ defined in (4). If the sources are uncorrelated, the mixture STFT coefficients $\mathbf{x}(n, f)$ also follow a zero-mean Gaussian distribution over this neighborhood with covariance

$$\mathbf{R}_{\mathbf{x}}(n, f) = \sum_{j=1}^{J} v_j(n, f)\, \mathbf{R}_j(f). \tag{5}$$

The log-likelihood of the data can then be defined as [4, 5]

$$\log \mathcal{L}_w = \sum_{n,f} \sum_{n',f'} w_{nf}^2(n', f') \log \mathcal{N}(\mathbf{x}(n', f')|0, \mathbf{R}_{\mathbf{x}}(n, f)) \tag{6}$$

where w_{nf} is a bi-dimensional window specifying the shape of the neighborhood such that $\sum_{n',f'} w_{nf}^2(n', f') = 1$. This model aims to improve the robustness of parameter estimation by locally exploiting the observed data in several T-F bins instead of a single one. Simple computation leads to

$$\log \mathcal{L}_w = -\sum_{n,f} \mathrm{tr}(\mathbf{R}_{\mathbf{x}}^{-1}(n, f)\widehat{\mathbf{R}}_{\mathbf{x}}(n, f)) + \log \det(\pi\, \mathbf{R}_{\mathbf{x}}(n, f)) \tag{7}$$

where $\mathrm{tr}(.)$ denotes the trace of a square matrix and $\widehat{\mathbf{R}}_{\mathbf{x}}(n, f)$ is the *local observed mixture covariance* in the neighborhood of (n, f) given by [4]

$$\widehat{\mathbf{R}}_{\mathbf{x}}(n, f) = \sum_{n',f'} w_{nf}^2(n', f')\mathbf{x}(n', f')\mathbf{x}^H(n', f'). \tag{8}$$

ML inference of the model parameters can now be achieved by maximizing (7), which does not require knowledge of $\mathbf{x}(n, f)$ anymore but of $\widehat{\mathbf{R}}_{\mathbf{x}}(n, f)$ only.

2.2 Auditory-motivated T-F representation

The ERB scale is an auditory motivated frequency scale defined by [7]

$$f_{ERB} = 9.26 \log(0.00437 f_{Hz} + 1). \tag{9}$$

For a given number of frequency bands, ERB-scale T-F representations exhibit higher frequency resolution than the STFT at low frequencies and lower resolution at high frequencies. We use the representation in [8] which differs from those used in CASA [7]

in two ways. Firstly, the bandwidth of each subband is not set to that of the cochlea but to a narrower value resulting in better separation performance. Secondly, the representation itself does not consist of a set of subband signals but of a set of local observed covariance matrices. Therefore, although it has never been employed in this context, it leads to straightforward computation of the ML criterion in (7).

The ERB-scale representation is computed as follows[1]. Let $H_f(t)$ be a bank of centered complex-valued bandpass filters associated with individual frequency bands f. These filters are defined as modulated Hanning windows. Their center frequencies are linearly spaced on the ERB scale between zero and the Nyquist frequency and the width of their main lobe is set to four times the difference between the central frequencies of adjacent filters. The mixture signal $\mathbf{x}(t)$ is split into subband signals $\mathbf{x}_f(t)$ defined by

$$\mathbf{x}_f(t) = \sum_{\tau} H_f(\tau)\mathbf{x}(t-\tau). \tag{10}$$

The local observed covariance is then computed from windowed subband signals by

$$\widehat{\mathbf{R}}_{\mathbf{x}}(n,f) = \sum_{t,f'} w_f^2(f') w_{\mathrm{ana}}^2(t-nL)\mathbf{x}_{f'}(t)\mathbf{x}_{f'}^H(t) \tag{11}$$

where $w_{\mathrm{ana}}(t)$ is an analysis window, $w_f(f')$ is a window specifying the shape of the frequency neighborhood and L the step between successive time frames.

3 Blind source separation algorithm

We now propose a source separation algorithm using the above ML criterion and applicable to ERB-scale time-frequency representations. This algorithm consists of three successive steps: initialization of the model parameters $v_j(n,f)$ and $\mathbf{R}_j(f)$ by hierarchical clustering and permutation alignment, ML parameter estimation by iterative expectation-maximization (EM), and derivation of the source images by Wiener filtering. We present the details of this algorithm for full-rank spatial covariances $\mathbf{R}_j(f)$ and briefly explain how to add the rank-1 constraint $\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f)$.

3.1 Parameter initialization and permutation alignment in the STFT domain

It has been shown that parameter initialization affects the source separation performance resulting from the EM algorithm [3]. In the case of a STFT representation, we follow the hierarchical clustering-based initialization [3] in which we have awared that the initialization robustness decreases when the number of active sources increases. We first normalize the vectors of mixture STFT coefficients so that they have unit Euclidean norm and the first entry has zero phase. In a given frequency bin f, each normalized vector of mixture STFT coefficients $\bar{\mathbf{x}}(n,f)$ is first considered as a cluster containing a

[1] Matlab software implementing this representation as well as ERB-scale multichannel filtering is available at http://www.irisa.fr/metiss/members/evincent/software.

single item. The distance between each pair of clusters is computed as the average distance between the associated normalized mixture STFT coefficients and the two clusters with the smallest distance are merged [3]. This linking process is repeated until the number of clusters is smaller than a fixed threshold K, which is much larger than the number of sources J [10], so as to eliminate outliers. We finally choose the J clusters C_j with the largest number of items denoted by $|C_j|$ and compute the associated spatial covariance matrices by

$$\mathbf{R}_j^{\mathrm{init}}(f) = \frac{1}{|C_j|} \sum_{\mathbf{x}(n,f) \in C_j} \mathbf{x}(n,f)\mathbf{x}^H(n,f). \tag{12}$$

The mixing vectors $\mathbf{h}_j^{\mathrm{init}}(f)$ are initialized in a similar fashion [3]. The order of $\mathbf{R}_j^{\mathrm{init}}(f)$ and $\mathbf{h}_j^{\mathrm{init}}(f)$ is then aligned across all frequency bins based on the DOA information carried by $\mathbf{h}_j^{\mathrm{init}}(f)$ [11] or by the first principal component of $\mathbf{R}_j^{\mathrm{init}}(f)$ [3] so as to correspond to the same source signal j. Finally, the source variances are simply initialized to $v_j^{\mathrm{init}}(n,f) = 1$.This basic initialization scheme performed as well as the more advanced but slower schemes in our experiment.

3.2 Parameter initialization and permutation alignment in the ERB domain

The above initialization procedure does not readily extend to ERB-scale representations for which T-F coefficients $\mathbf{x}(n,f)$ are not available. While this procedure can be applied to cluster the first principal components of the local observed covariances $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$, we found that this did not result in good performance in the high frequency range. Indeed, due to broadness of high frequency subbands, the reduction of local observed covariances to their first principal components results in some information loss.

In order to ensure a fair comparison of both representations independently of the parameter initialization procedure, we derive the initial parameters in the ERB domain from those in the STFT domain as follows. Given the estimated clusters C_j, initial estimates of the source images in the STFT domain are obtained by binary masking as

$$\mathbf{c}_j^{\mathrm{init}}(n,f) = \begin{cases} \mathbf{x}(n,f) & \text{if } \bar{\mathbf{x}}(n,f) \in C_j \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{13}$$

The order of $\mathbf{c}_j^{\mathrm{init}}(n,f)$ is aligned across frequency based on $\mathbf{h}_j^{\mathrm{init}}(f)$ as above. ERB-domain initial local observed covariances $\widehat{\mathbf{R}}_{\mathbf{c}_j}^{\mathrm{init}}(n,f)$ are then derived by STFT inversion of $\mathbf{c}_j^{\mathrm{init}}(n,f)$ followed by ERB-scale representation. Finally the spatial covariance matrices $\mathbf{R}_j^{\mathrm{init}}(f)$ are initialized by averaging $\widehat{\mathbf{R}}_{\mathbf{c}_j}^{\mathrm{init}}(n,f)$ over all time frames n, while the initial mixing vectors $\mathbf{h}_j^{\mathrm{init}}(f)$ are taken as the first principal component of $\mathbf{R}_j^{\mathrm{init}}(f)$. Again, the source variances are initialized to $v_j^{\mathrm{init}}(n,f) = 1$.

3.3 ML parameter estimation by EM

We now wish to optimize the ML criterion in (7). Since this criterion does not assume a particular representation, a single parameter estimation algorithm can be designed both

the the STFT and the ERB-scale representation. The EM algorithm is well known as an appropriate choice in this case. We reuse the M-step of the EM algorithm derived for rank-1 and full-rank covariance models in [3] but adapt the E-step to account for the use of the local observed covariance. Note that, strictly speaking, this algorithm and that in [3] are generalized forms of EM since the M-step increases but does not maximize the likelihood of the complete data.

For the full-rank model, we consider the set of T-F coefficients of all source images on all channels and all time frames $\{\mathbf{c}_j(n,f)\ \forall j,n\}$ as the EM *complete data*. The details of one EM iteration for each source j in frequency bin f are as follows. In the E-step, the Wiener filter $\mathbf{W}_j(n,f)$ and the covariance matrix $\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$ of the estimated source image are computed as

$$\mathbf{W}_j(n,f) = \mathbf{R}_{\mathbf{c}_j}(n,f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f) \tag{14}$$

$$\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f) = \mathbf{W}_j(n,f)\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)\mathbf{W}_j^H(n,f) + (\mathbf{I} - \mathbf{W}_j(n,f))\mathbf{R}_{\mathbf{c}_j}(n,f) \tag{15}$$

where $\mathbf{I}$ is the $I \times I$ identity matrix, $\mathbf{R}_{\mathbf{c}_j}(n,f)$ is defined in (4), $\mathbf{R}_{\mathbf{x}}(n,f)$ in (5) and $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$ in (8). In the M-step, $\mathbf{R}_j(f)$ and $v_j(n,f)$ are updated as [3]

$$v_j(n,f) = \frac{1}{I}\,\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)) \tag{16}$$

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(n,f)}\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f) \tag{17}$$

Note that, compared to the EM algorithm in [3], $\widehat{\mathbf{R}}_{\mathbf{c}_j}(n,f)$ is computed directly from $\widehat{\mathbf{R}}_{\mathbf{x}}(n,f)$ in the E-step instead of the estimated source images.

This modification of the E-step is also applied to the rank-1 model. Due to the lack of space we do not describe the details of the EM algorithm for the rank-1 model here. EM improves the separation performance compared to the initialization alone [3].

3.4 Source separation by Wiener filtering

The spatial images of all sources can be estimated given $v_j(n,f)$ and $\mathbf{R}_j(f)$ by Wiener filtering. In the STFT domain, this filter is applied to the mixture STFT coefficients as

$$\widehat{\mathbf{c}}_j(n,f) = v_j(n,f)\mathbf{R}_j(f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f) \tag{18}$$

and time-domain signals are obtained by STFT inversion. In the ERB domain, this filter is applied to the windowed subband signals $w_{\mathrm{ana}}(t-nL)\mathbf{x}_f(t)$ instead. Subband source image signals are obtained by overlap-and-add via

$$\widehat{\mathbf{c}}_{jf}(t) = \sum_n w_{\mathrm{syn}}(t-nL)w_{\mathrm{ana}}(t-nL)v_j(n,f)\mathbf{R}_j(f)\mathbf{R}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}_f(t) \tag{19}$$

where $w_{\mathrm{syn}}(t)$ is a synthesis window satisfying $\sum_t w_{\mathrm{syn}}(t-nL)w_{\mathrm{ana}}(t-nL) = 1$. Fullband signals are finally recovered by least-squares filterbank inversion [8].

4 Experimental results

We compared the performance of the blind source separation algorithms presented in this paper and those in [3], based either on a rank-1 or a full-rank model, on the use of the STFT or the ERB-scale representation as input, and on the choice of the ML criterion based on the mixture STFT coefficients or on the local observed covariance (LOC). We generated three stereo synthetic mixtures of three speech sources by convolving different sets of 10 s speech signals sampled at 16 kHz with room impulse responses simulated via the source image method for a room with reverberation time $T_{60} = 250$ ms. The distance between the microphones was 5 cm and that from the sources to microphones was 50 cm. The STFT was computed with half-overlapping sine windows of length 1024 and the bi-directional window w defining the shape of neighborhood was the outer product of two Hanning windows of length 3. The ERB-scale representation was computed with a bank of 250 filters and half-overlapping sine analysis and synthesis windows, *i.e.* w_{ana} and w_{syn}, of length 1024. The number of EM iterations for all algorithms was 10. Note that the optimal number of frequency bin for the STFT and the ERB-scale representation are different. Binary masking [1] and ℓ_0-norm minimization [2] were also evaluated as baseline approaches with the same mixing vectors used for the rank-1 model. Separation performance was evaluated using the signal-to-distortion ratio (SDR) criterion measuring overall distortion, as well as the signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) and source image-to-spatial distortion ratio (ISR) criteria in [12], averaged over all sources and all mixtures.

The results in Table 1 indicate that the full-rank model always provides higher SDR and SAR than both the rank-1 model and baseline approaches which comforts our claim that it better approximates the reverberant mixing process [3], even with other representations than the STFT. The exploitation of local observed covariance is shown to be beneficial for both rank-1 and full-rank models since it provides 0.3 dB SDR improvement in both cases and also increases all other performance criteria, especially the SAR. The ERB-scale T-F representation provides an additional 0.5 dB SDR enhancement compared to sole use of the local observed covariance for the full-rank model. However, this representation results instead in a decrease of the SDR when the rank-1 model is used. This suggests that full-rank models are needed from now to achieve further improvements in ERB scale-based source separation, since they partially overcome the narrowband assumption and do not suffer from the large bandwidth at high frequencies.

In the end, the combination of local observed covariance and ERB-scale representation improves the SDR, SIR, SAR, and ISR by 0.8 dB, 1.4 dB, 0.7 dB, and 1 dB, respectively compared to the full-rank model based algorithm proposed in [3] and outperforms all other approaches. This improvement satisfies our theoretical expectations and even appears more positively when considering the fact that the new algorithm is faster due to the processing of 250 ERB-scale frequency bins instead of 513 STFT bins.

5 Conclusion

In this paper, we presented an under-determined convolutive source separation algorithm offered by the combination of local observed covariance, auditory-motivated T-F representation and the full-rank covariance model. We addressed the estimation of

Approach	SDR	SIR	SAR	ISR
Full-rank model + LOC + ERB	6.6	10.1	9.3	10.9
Full-rank model + LOC	6.1	8.9	9.0	10.2
Full-rank model	5.8	8.7	8.6	9.9
Rank-1 model + LOC + ERB	4.1	7.7	8.4	8.9
Rank-1 model + LOC	4.6	8.1	8.5	9.2
Rank-1 model	4.3	7.9	7.9	9.2
Binary masking	4.8	10.2	5.6	10.3
ℓ_0-norm minimization	3.8	7.9	6.9	9.2

Table 1. Average source separation performance (dB) of stereo mixtures of 3 speech sources

.

model parameters by EM algorithm with a proper parameter initialization scheme. Experimental results over speech mixtures confirm that both local covariance model and auditory-motivated T-F representation are beneficial for full-rank model based source separation. Consequently, the proposed approach outperforms both the full-rank model and rank-1 model in [3] and baseline approaches in a reverberant environment.

References

1. Yılmaz, O., Rickard, S.T.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. on Signal Processing **52**(7) (July 2004) 1830–1847
2. Vincent, E., Jafari, M.G., Abdallah, S.A., Plumbley, M.D., Davies, M.E.: Probabilistic modeling paradigms for audio source separation. In Wang, W., ed.: Machine Audition: Principles, Algorithms and Systems. IGI Global to appear.
3. Duong, N.Q.K., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Trans. on Audio, Speech and Language Processing (2010) To appear.
4. Vincent, E., Arberet, S., Gribonval, R.: Underdetermined instantaneous audio source separation via local Gaussian modeling. In: Proc. ICA. (2009) 775–782
5. Févotte, C., Cardoso, J.F.: Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In: Proc. WASPAA. (2005) 78–81
6. Deville, Y.: Temporal and time-frequency correlation-based blind source separation methods. In: Proc. ICA. (2003) 1059–1064
7. Roman, N., Wang, D., Brown, G.: Speech segregation based on sound localization. Journal of the ASA **114**(4) (October 2003) 2236–2252
8. Vincent, E.: Musical source separation using time-frequency source priors. IEEE Trans. on Audio, Speech and Language Processing **14 (1)** (2006) 91–98
9. Burred, J., Sikora, T.: Comparison of frequency-warped representations for source separation of stereo mixtures. In: Proc. 121st AES Convention. (October 2006)
10. Winter, S., Kellermann, W., Sawada, H., Makino, S.: MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization. EURASIP Journal on Advances in Signal Processing **2007** (2007) article ID 24717.
11. Sawada, H., Araki, S., Mukai, R., Makino, S.: Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. IEEE Trans. on Audio, Speech, and Language Processing **15**(5) (July 2007) 1592–1604
12. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.: First stereo audio source separation evaluation campaign: data, algorithms and results. In: Proc. ICA. (2007) 552–559